Multi-Animal Tracker Leveraging Comprehensive Visual Features

Sewon Lee Daseul Park Hyeonseo Hwang Kwangeun Yeo
Seoul National University, South Korea
{sewon0803, dspark12, vadanamu, kwangeun.yeo}@snu.ac.kr

Abstract

Multi-animal tracking (MAT), a type of Multi-object tracking (MOT), is a challenging task in computer vision that has great importance in various fields, including neuroscience, wildlife conservation, and ecology. However, MAT remains largely unexplored, and the unique characteristics in motions and appearance similarities of animals, resulting in frequent identity switches, make this task even more challenging. To address these challenges, we propose a novel approach that goes beyond traditional reliance on 2D motion and high-level visual cues. Our approach incorporates comprehensive visual features for robust tracking, specifically leveraging depth information and multi-level appearance cues to tackle the notorious identity switch problems in MAT. By utilizing depth information, one can reframe the 2D tracking problem into 3D framework, thereby enhancing the differentiation, especially after occlusions. Multi-level appearance cues are the fused low-level and high-level feature representations, which can capture more details of each animal. Built upon the DeepSORT algorithm, our proposed model significantly improves the performance of the baseline, especially in the association task, shown by the increase in IDF1 score. Next, we investigate the contributions of each proposed visual feature—depth and multi-level appearance feature— by conducting ablation study. We further dissect the effects of each feature by analyzing each feature's effect in sequence-level, which shows that these features —depth in particular— contribute a lot in improving association quality of the tracker.

1. Introduction

Multi-object tracking (MOT) is a popular yet challenging task in computer vision. It aims to locate multiple objects of interest in a video sequence and generate each object's trajectory, while maintaining their identities. Particularly, multi-animal tracking (MAT), a kind of MOT, is critical for animal motion and behavioral analysis, and thus has great importance in neuroscience, wildlife conservation and ecology, to name a few.

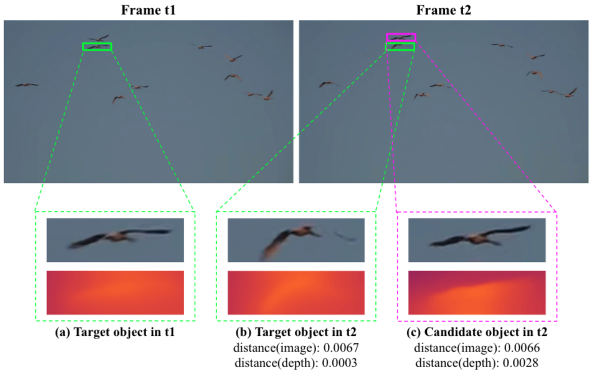


Figure 1. The distance between visual features of the objects in two consecutive frames

Despite its importance, MAT is highly unexplored. Most MOT studies have focused on tracking pedestrians or vehicles [10, 13, 31, 60]. However, animals have certain characteristics in their motions and appearance patterns that are significantly different from humans or vehicles, making the task more challenging [55]. While humans and vehicles have various visual cues to distinguish them (*e.g.*, colors and shapes), animals have extremely high visual similarity to each other. This makes it difficult to rely solely on these appearance cues for tracking, which often results in identity switches. The pose variations (*e.g.*, flying, swimming and walking) and resulting complexity in motion patterns also pose a challenge in detecting and tracking animals. Indeed, most trackers targeting pedestrians or vehicles show significant performance drop when tested on animal tracking dataset from [55]. It raises the need to propose a novel approach, dedicated to tracking multiple animals.

Previous approaches in MOT, let alone MAT, have primarily relied on 2D motion information and/or appearance cues [2,4,6,51,58]. More precisely, motion information includes the x, y coordinates and velocity of detected bounding boxes. Appearance cues are usually represented as high-level visual features of each object, and are commonly used in re-identification. Nonetheless, these approaches fail to offer reliable tracking in multi-animal scenarios due to challenges such as high visual similarity, irregular motions, frequent interactions, and resulting occlusions.

To address this challenge, we propose a novel approach that incorporates comprehensive visual features for robust tracking, moving beyond the conventional reliance solely on 2D motion and high-level visual cues. Our first proposed visual feature is image depth. In Figure 1, we illustrate the difficulty of distinguishing the target object from a candidate object using high-level image feature distances, as these distances are nearly identical (0.0067 and 0.0066). However, by leveraging depth information, we can distinctly identify the target object, which exhibits a distance ten times closer to that of the candidate object. This suggests that expanding the 2D-context to 3D has the potential to solve the occlusion problem, which is common in MAT. To further enhance the model's ability to identify each animal, we seek to capture each individual's appearance cues in more detail, by utilizing low-level features along with high-level features. This approach is expected to be especially helpful for tracking animals, given their extremely high visual similarity and the common failure to capture the identity only with high-level features [55]. Finally, we conduct ablation studies to investigate the contribution of each proposed visual feature in MAT performance.

Our contributions can be summarized as follows:

- We propose a novel approach for MAT, which is to leverage a depth module. To the best of our knowledge, it is the very first attempt to directly utilize depth information extracted from 2D videos within the context of MAT.

- We further propose to utilize multi-level appearance representation, which is expected to solve the homogeneous appearance problem in animals.

- We analyze the effect of different features –depth-incorporated motion, and multi-level appearance cues– in tracking animals.

We hope these novel approaches and following analysis to facilitate meaningful discussions in the under-explored field of MAT.

2. Related Works

2.1. Multi-Object Tracking Algorithms

Tracking-by-Detection (TD) is a predominant paradigm in MOT, which decomposes the MOT task into two subtasks: object detection and data association. In the object detection task, [23,43] instances are detected in each frame using pretrained detectors such as YOLO [41]. In the data association task, the detected instances are associated with object identity to generate trajectories using optimization techniques (*e.g*., Hungarian algorithm [4] and network flow algorithm [9]).

SORT [4] is an online TD MOT framework that utilizes the Kalman filter [19] for velocity model estimation and the Hungarian algorithm for data association, and this design has been widely adopted in recent MOT models. DeepSORT [51] mitigated the identity switching and occlusion issue of SORT by using nearest-neighbor queries and appearance feature representations. OC-SORT [7] harnessed observation-centric re-update for reducing the accumulation of errors. QDTrack [35] aimed to enhance data matching by using contrastive learning to learn the similarity between instances. ByteTrack [56] achieved state-of-the-art performance by recovering true object identities from low-score bounding boxes instead of neglecting them.

There are multiple paradigms other than TD in MOT. Tracking-by-segmentation (TS) leverages image segmentation models such as Mask R-CNN [14] to utilize pixel-level information of instances. MOTS [46] has demonstrated that MOT and image segmentation tasks can be effectively coupled. Utilizing an explicit store of memory for segmentation has been suggested in space-time memory (STM) [33]. By extracting memory embedding from each frame, features that exist over multiple frames could be effectively utilized for tracking. This idea has been expanded in XMem [8] by using multiple memory stores inspired by the neurological model of memory. Recently, Track Anything framework [52] showed that using segment anything model (SAM) [20] in conjunction with XMem can achieve robust performance.

Besides TD and TS, tracking-by-regression (TR) and tracking-by-attention (TA) paradigms have been suggested. Tracktor [2], which is a TR model, used an object detector with a bounding box regression head and an object classification head, which efficiently handles the occlusion of tracks. CenterTrack [59] is a simple TR model that represents objects as points and associate the instances using greedy matching. Trackformer [29] is a TA model that utilized transformer encoder-decoder architecture for MOT.

2.2. Multi-Animal Tracking Algorithms

For MAT, several models based on existing MOT algorithms have been suggested. DeepLabCut [28] is an extension of the multi-human pose estimation model DeeperCut [17], which is based on ResNet [15]. While the DeepLabCut model showed state-of-the-art performance in animal body part tracking, it only had limited capability for MAT. IDtracker.ai [45] is a CNN-based MAT framework that utilizes a specialized crossing detector to segment videos into non-occluded chunks with respect to each instance. TRex [47] is a TS-based MAT model that achieved performance comparable to IDtracker.ai with improvement in the pro-

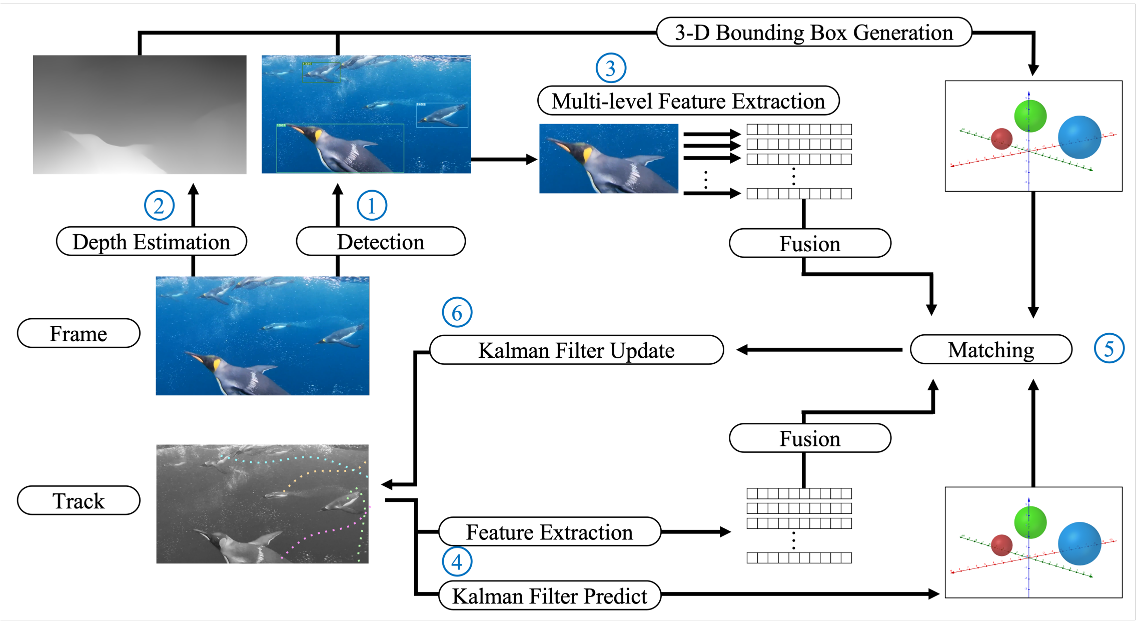


Figure 2. Architecture of the proposed MAT model leveraging comprehensive visual features.

cessing speed. SLEAP [36] is a general multi-animal pose tracking framework that offers both visual feature-based tracking and flow-shift tracking.

2.3. Object Tracking Using Depth

Traditional object tracking models [2,4,51] that depend only on RGB image decreased accuracy when distinguishing objects with similar appearances, such as objects with similar textures. To address this limitation, algorithms have been suggested that utilize both RGB images and depth information.

VGG3D using RGB-D [61] proposed a method to improve object recognition accuracy by including 3D shape and distance information of objects using deep convolutional neural networks (DCNN) trained on RGB-D images. DAL [37] proposed a method that utilizes depth information in the field of Long-term Tracking. The results achieved high tracking accuracy and stability compared to previous Long-term Tracking methods. And the algorithm also performed well on large-scale datasets and was sensitive to object size and distance. Some works focused on reducing occlusion by integrating depth with MOT. 3DT [16] and AB3DMOT [50] aimed to resolve the occlusion problem in MOT by using depth information to generate 3D trajectories.

In order to utilize these methods, depth information, obtained through specialized devices such as RGB-D, LiDAR, or GPS, is required. However, these types of data are not generally available for MAT tasks. This issue can be mitigated by extracting depth information from RGB images using depth estimation models. DET [24] and DP-MOT [38] are pioneering works that integrated depth estimation from RGB images with conventional TD-based MOT.

Depth estimation is a common task in computer vision where the goal is to predict the distance of objects or scenes from one or more images. There are three types of depth estimation: monocular (using one image), binocular (using two images), and multi-view (using more than two images) [27]. For our research, we will be focusing on monocular depth estimation (MDE) methods because they can predict a wider range of depth [1, 34]. Deep learning-based methods have become highly effective for monocular depth estimation, demonstrating superior performance compared to other approaches. These methods can be broadly classified into two categories based on their underlying architecture: convolutional neural network (CNN)-based [11, 22, 32, 53] and Transformer-based [5, 39, 54] methods. In recent years, researchers have explored more advanced architectures [5, 11, 12, 12, 30] to further improve the accuracy of monocular depth estimation, demonstrating promising results.

3. Method

Following the tracking-by-detection paradigm, the proposed method combines depth-incorporated object positions and multi-level appearance feature representations of detected objects. The extracted features are then used in association step, when computing the similarity between current detections and the estimated state of existing objects in current frame. The proposed pipeline in illustrated in

Fig. 2. In the following sections, the method is described in detail by its main components, 1) detection, 2) depth-incorporated track handling, and 3) multi-level appearance feature extraction and association.

3.1. Detection

To locate animals in each incoming frame, we utilize YOLOv5 [18]. YOLOv5, similar to other detectors, consists of three key components. The first component is a CNN-based backbone that extracts image features. The second component consists of the neck layers, which combine and integrate image features before forwarding them for prediction. The final component is the detection head, responsible for predicting object classes and bounding boxes. YOLOv5 utilizes the architecture of CSPDarknet53 [48] with an SPP layer serving as the backbone, PANet [25] as the neck layers, and a YOLO detection head [42]. Due to its notable reputation and user-friendly nature as a one-stage detector, we select YOLOv5 as our detection module.

3.2. 3D Tracking

3.2.1 Depth Estimation

One of the main ideas of the proposed method is to leverage depth value for each detected object (*i.e.*, extending the 2D image space into 3D) (Fig. 2 ②). While occluded objects are likely to share the same 2D coordinates, the 3D coordinates are guaranteed to be unique for each object. Therefore, we expect this approach to solve the occlusion problem, which is frequent in multi-animal settings.

To get the depth value of each detection, we utilized pre-trained MiDaS (dpt-beit-large-512) [40] as our depth estimation model. This model is pre-trained on 12 datasets (ReDWeb, DIML, Movies, MegaDepth, WSVD, TartanAir, HRWSI, ApolloScape, BlendedMVS, IRS, KITTI, NYU Depth V2) of several depths and environments. After obtaining the entire depth map of a given frame with MiDaS, the bounding box region of each object is cropped from the depth pixels.

The depth map of each bounding box is divided into 9 patches for each detection, and only the central patch is used for depth estimation. This is to prevent the background pixels affecting the z coordinate of the bounding boxes. The center z coordinate of each bounding box is calculated by the median, and the length along the z-axis was calculated by the IQR of depth values. Combined with the 2D coordinates obtained from the detection step, the z-axis coordinates are used to represent the 3D coordinates of each detected animal.

3.2.2 3D State Estimation

For state estimation (Fig. 2 ④), we adopt a 3D Kalman filter based on [19] to predict the 3D coordinates of existing objects in current frame t, based on frame $t-1$. Kalman filter uses a linear constant velocity model, based on coordinates from the previous frames. The standard deviation of the positions and velocities were scaled by the height of the respective bounding box. We assumed that the z-axis motion is more nonlinear than the x-y plane motion, and thus scaled the z-axis standard deviation by a factor of 4. For IoU calculation, the 3D bounding boxes were projected to the x-y, y-z, and x-z planes, and the IoU was calculated on each plane. 3D IoU was obtained by calculating the weighted mean of 2D IoUs on the three planes.

3.3. Association

Association is done by computing the similarity between the latest detections and the estimated state of existing objects (*i.e.*, tracks), and assigning each detection with the closest object. Building upon the DeepSORT [51] algorithm, we employ a comprehensive approach that combines both motion and appearance information to effectively address the assignment problem. Moreover, we employ a novel multi-level appearance feature extractor that utilizes low-level features, enabling the capture of finer details in images.

3.3.1 Multi-Level Appearance Feature Extractor

We utilize a convolutional neural network (CNN) to extract appearance features. CNNs are designed to better comprehend the overall semantics of an image as the receptive field expands. As a result, high-level features in CNNs typically reflect abstract semantics, while low-level features capture finer details. To address the significant visual similarities among animals, we employ a multi-level feature extractor that utilizes detailed information.

The overall architecture is illustrated in Figure 3. It begins with a 3×3 convolutional layer, which transforms the input image into concise high-dimensional feature maps. These feature maps then undergo processing through four consecutive convolutional blocks. Each block consists of a 3×3 convolution, followed by a ReLU activation function, and another 3×3 convolution. This hierarchical arrangement enables the extraction of features at different levels. Throughout this process, the channel dimensions of the feature maps increase, while the spatial resolution remains unchanged.

Next, we combine the four multi-level features using fusion operations. Three different fusion operations are experimented. The first function is a simple addition, and the second one is concatenation followed by a fully connected layer. The last operation is a gate fusion that computes a weighted average of features using trainable coefficients:

$$\sigma(FW^F)F \tag{1}$$

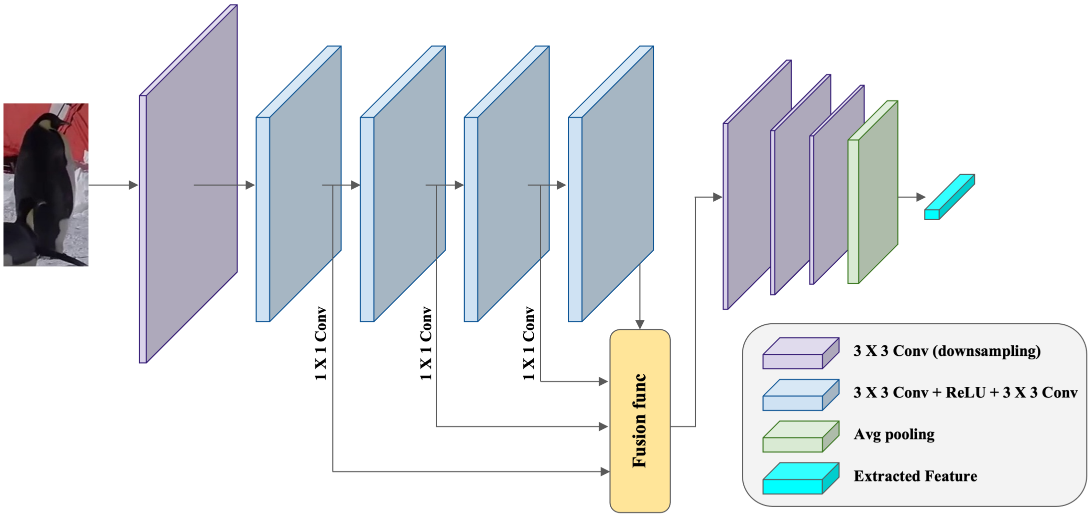


Figure 3. Architecture of the multi-level appearance feature extractor

where F represents the matrix form of the multi-level features, and W^F denotes a trainable parameter that maps the features to one-dimensional gate vectors.

Following the fusion operation, additional convolutional layers and an average pooling layer are added to extract the features for the the tracking algorithm. Then the network is trained on the ID classification task, which involves classifying the individual IDs of the provided animal images. To facilitate this task, additional dense layers are introduced. These layers are only employed during the training phase and are not activated in the tracking algorithm.

3.3.2 Assignment

To assign the detected animal to a proper object, we calculate the weighted sum of distances (c_{ij}) of each feature embedding (Fig. 2 ⑤):

$$c_{ij} = \lambda d_{ij}^{(1)} + (1 - \lambda) d_{ij}^{(2)}. \tag{2}$$

Here, d_{ij} represents the distance between 3D coordinates ($d^{(1)}$) and appearance feature ($d^{(2)}$) of i-th detected object and the estimated state of j-th existing object for the current frame, respectively.

Based on the similarity score, the Hungarian algorithm [21] assigns each detected object to an existing track. Once the match is done, Kalman filter is updated with the newly associated object (Fig. 2 ⑥).

4. Experiments

Within this section, we assess the efficacy of the proposed method by presenting a comprehensive analysis of experimental results. The conducted experiments serve to address a set of research questions (RQs):

- **RQ1**: How does the performance of the proposed method compare to baseline approaches?
- **RQ2**: What impact does the utilization of each comprehensive visual feature has on the results?
- **RQ3**: How does the performance of the method vary across different animal species?

4.1. Experimental Settings

4.1.1 Dataset

AnimalTrack AnimalTrack [55] dataset is a fully annotated video dataset dedicated to MAT in the wild. It contains 32 videos (11,500 images) for training and 26 videos (13,200 images) for testing. There are total 10 animal categories and each video contains only one category of animal. The video frame rate is 30 FPS and every frame is annotated. The annotation includes animal trajectory ID, coordinates, size and visibility ratio of the detected animals.

4.1.2 Evaluation

To evaluate the tracking performance, we use multiple metrics including commonly used CLEAR MOT [3] metrics and more recently proposed HOTA (higher order tracking accuracy) [26] metrics.

HOTA [26] is a commonly used metric that takes into account both detection and association quality. This makes HOTA a balanced metric compared to MOTA (multi-object

Tracker	HOTA	MOTA	IDF1	IDP	IDR	IDSW↓	MT	PT	ML↓	FP↓	FN↓	FM↓
JDE [49]	26.8	27.3	31.0	51.0	22.0	3,187	106	414	584	17,887	155,623	5,031
FairMOT [57]	30.6	29.0	38.8	62.8	28.0	2,335	143	462	499	17,653	152,624	5,447
CenterTrack [58]	9.9	1.6	7.0	8.9	5.8	89,655	265	423	416	32,050	117,614	7,583
Tracktor++ [2]	44.2	55.2	51.0	58.5	45.1	1,976	364	472	268	25,477	81,538	4,149
ByteTrack [56]	40.1	38.5	51.2	64.9	42.3	1,309	310	465	329	31,591	116,587	3,513
Trackformer [29]	31.0	20.4	36.5	40.9	32.8	4,355	230	491	383	70,404	118,724	3,725
DeepSORT [51]	32.8	41.4	35.2	49.7	27.2	3,503	213	452	439	14,131	124,747	4,527
Ours	32.5	41.0	37.4	41.6	34.0	6,215	383	432	295	46,417	90,892	5,360

Table 1. Evaluation results of different trackers on AnimalTrack

Multi-level	Depth	HOTA	MOTA	IDF1	IDP	IDR	IDSW↓	MT	PT	ML↓	FP↓	FN↓	FM↓
-	-	32.4%	41.1%	37.0%	41.4%	33.5%	6,345	377	440	293	**45,109**	91,811	**5,319**
✓	-	32.5%	**41.1%**	37.4%	**41.9%**	33.8%	6,351	371	**447**	**292**	45,122	91,799	5,337
-	✓	32.4%	40.9%	**37.6%**	41.7%	**34.1%**	6,275	376	442	**292**	46,537	**90,816**	5,371
✓	✓	**32.5%**	41.0%	37.4%	41.6%	34.0%	**6,215**	**383**	432	295	46,417	90,892	5,360

Table 2. Evaluation results of ablation and final models

tracking accuracy) [3] and IDF1 [44], which are also commonly used but more biased towards detection and association quality, respectively.

MOTA [3] takes into account FP (false positives), FN (false negatives), and IDSW (ID switches). FP and FN indicates the total number of false-positive and -negative tracks, and IDSW measures the total number of identity switches across the entire dataset.

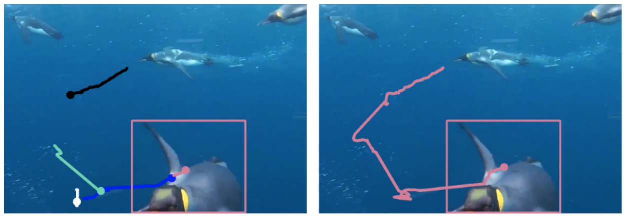
Identification (ID) metrics from [44] are also the main metrics of our interest. They include IDF1, IDP (ID precision), and IDR (ID recall). IDF1 is calculated as the ratio of correct detections to the average of ground-truth and predicted detections.

Since our approach primarily aims to increase association quality in MAT, we focus on HOTA and ID metrics, including IDF1 and IDSW.

4.2. Overall Comparison (RQ1)

4.2.1 Baseline Models

To see how existing tracking models perform on multi-animal dataset, we evaluated 7 widely-used trackers and set their results as baselines. The included trackers are DeepSORT [51], Tracktor++ [2], JDE [49], CenterTrack [58], FairMOT [57], ByteTrack [56], and Trackformer [29], all the way from classic models to more recent transformer-based models. Each model is trained on AnimalTrack train set with its default architecture, including detector, and evaluated on the test set.



Baseline

Ours

Figure 4. Example track produced by different trackers

4.2.2 Experimental Results

The overall performance of different tracking algorithms are summarized in Table 1 [55]. The results reveal the superiority of modern models like Tracktor++ [2] and ByteTrack [56] over other models across multiple dimensions. Conversely, models based on simpler approaches like our method and DeepSORT [51] exhibit relatively inferior performance.

However, it is important to highlight that our method surpasses the majority of baseline models in the key metric we focused on, specifically IDF1, which directly addresses the ID switch problem. Notably, our method demonstrates superior performance compared to the target model DeepSORT [51] in effectively handling the ID switch problem. This can be observed in Figure 4, where the line colors represent distinct track IDs. In contrast to baseline models, which frequently experience ID changes within video frames, our method ensures consistent tracking without any ID switches.

Furthermore, our method exhibits model-agnostic characteristics, making it compatible with a wide range of trackers that utilize motion and appearance information. The observed superiority of our method over the target model substantiates the potential improvement that can be achieved by integrating our method into other tracking frameworks. Future research endeavors may explore the integration of comprehensive visual features into alternative models, leaving room for further advancements in this area.

4.3. Improvement Analysis (RQ2)

Ablation study The performance of our final and ablation models are shown in Table 2. The results show that the key features of our model, utilization of depth feature and multi-level appearance feature, contribute to robust performance. Moreover, our final model significantly outperforms DeepSORT (Table 1) in most metrics. Specifically, our model showed IDF1 2.2%p higher than DeepSORT, which demonstrates the robust association capability of our model. More detailed analysis of model performance in each animal species and video is provided in section 4.4.

Performances w.r.t. Depth Features Our multi-animal tracker utilizes the depth estimation network to construct the track of each animal in 3-D space. As shown in Table 2, the integration of the depth feature improved the tracking performance. Specifically, IDF1 was increased by 0.5%p by adding the depth feature to the base model. This validated our hypothesis that depth information can associate bounding boxes in a situation where 2-D state and appearance embedding cannot provide enough information for the association.

Performances w.r.t. Multi-level Appearance Features We propose three distinct operations for the fusion of multi-level features: summation, concatenation, and a gate mechanism. Table 3 reveals that all fusion methods outperform the base DeepSORT [51] method. While the differences in performance may not be significant, these findings underscore the potential efficacy of low-level fusion. Notably, with regard to ID-related metrics, all fusion methods demonstrate enhanced performance compared to the base methods. Among the fusion methods, simple summation yields the most favorable results. Conversely, the concatanation and gate methods, incorporating trainable parameters, exhibit relatively inferior performance. This observation can be attributed to the relatively simpler patterns involved in the integration of low-level and high-level features. The methods employed for feature extraction and combination at multiple levels hold promise for achieving further improvements, although such exploration is deferred as a topic for future research.

Method	HOTA	MOTA	IDF1	IDP	IDR
BASE	32.38	41.08	37.04	41.44	33.48
SUM	**32.51**	41.08	**37.41**	**41.85**	**33.82**
CONCAT	32.20	**41.18**	37.25	41.67	33.68
GATE	32.36	41.08	37.38	41.81	33.80

Table 3. Performance Comparison of Multi-Level Feature Fusion Methods

4.4. Qualitative Analysis (RQ3)

To investigate the effects of proposed methods in a more detailed manner, we compare model performances on each video sequence. The list of the analyzed models is same with the one from ablation study (Sec. 4.3.). Figure 5 shows the model HOTA, MOTA, and IDF1 scores on representative sequences, where there are significant differences in model performance. Based on HOTA scores, each model performance varies a lot between sequences, and there is no dominant winner in general. While the proposed method outperforms the baseline model in some sequences (*e.g*., zebra2 and penguin2), the baseline model also outperforms the proposed model in other sequences, such as deer3. MOTA scores are generally comparable between models, except for deer1. This may be due to the fact that MOTA is a metric that is more biased towards detection than association. Since all models shared the same detector (fine-tuned YOLOv5), the detection performance would be similar across all models, leading to similar MOTA scores. Comparison of IDF1 scores reveals that the proposed model with depth and multi-level features shows better performance than other models in general, and depth information especially contribute a lot in increasing IDF1 score. This indicates that utilizing depth information and multi-level appearance features can improve the tracker performance, especially in association.

Taking a deeper dive, we further investigate the cause of contrasting effects of proposed methods on different video sequences. Taking deer3 and zebra2 as examples, we notice notable differences in camera motion or magnification ratio that can be an external cause of depth value change. While zebra2 sequence only has a mild left/right movement (*i.e*., movement along the x and y axis), deer3 sequence shows a significant camera motion, such as rotation and zooming in. These external changes (*i.e*., changes that do not originate from the animal movement itself) may cause confusion in the proposed model that utilizes depth information for tracking. Therefore, it is expected that our proposed model would gain a significant performance boost after compensating for these camera motions–especially in z-axis movements– and magnifications, which is a topic for future research.

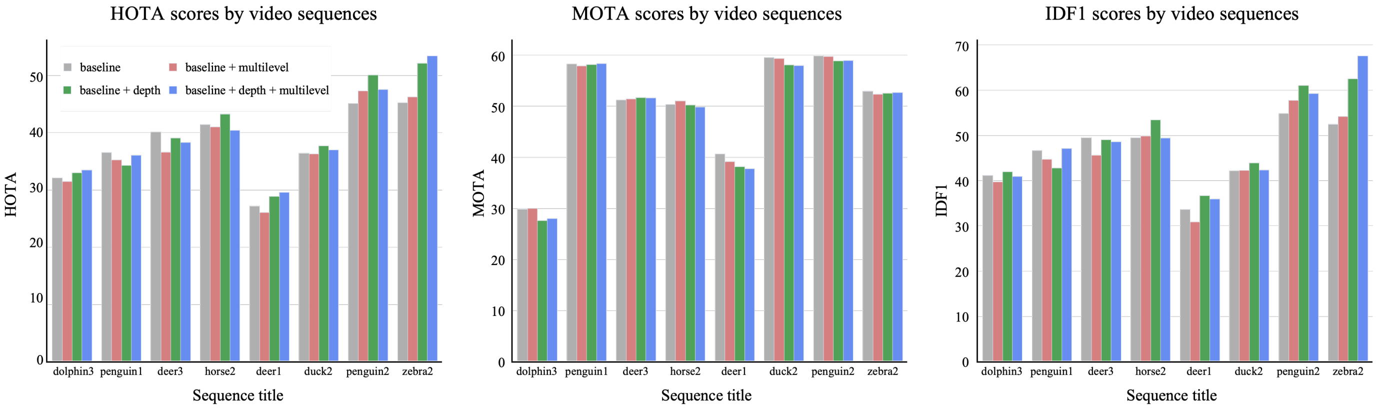


Figure 5. Sequence-level analysis of model performance

5. Conclusion

Our work pioneered the utilization of depth estimation and multi-level visual feature extraction for multi-animal tracking. Importantly, we show that both depth estimation and multi-level feature extraction independently contribute to robust tracking performance, enhancing the identity consistency. Although the limited capability of the base model, DeepSORT, prevented us from achieving the state-of-the-art performance, we reason that our two main contributions can be integrated with any tracking model, including ByteTrack and Tracktor++, to increase the tracking performance. Specifically, we show that our techniques significantly improve the association performance (IDF1) of the tracking model, which is the main challenge in MAT tasks. Hence, we speculate that our findings can be utilized to build a state-of-the-art MAT model, which would be a reliable and valuable asset for biomedical researchers.

References

[1] Robert S Allison, Barbara J Gillam, and Elia Vecellio. Binocular depth discrimination and estimation beyond interaction space. *Journal of Vision*, 2009. 3

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 1, 2, 3, 6

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5, 6

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016. 1, 2, 3

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[6] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017. 1

[7] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking, 2023. 2

[8] Ho Kei Cheng and Alexander G. Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model, 2022. 2

[9] Afshin Dehghan, Yicong Tian, Philip HS Torr, and Mubarak Shah. Target identity-aware network flow for online multiple target tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015. 2

[10] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 2014. 3

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[16] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krähenbühl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. *CoRR*, abs/1811.10742, 2018. 3

[17] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model, 2016. 2

[18] Glenn Jocher, Alex Stoken, Jirka Borovec, Ayush Chaurasia, Liu Changyu, Adam Hogan, Jan Hajek, Laurentiu Diaconu, Yonghye Kwon, Yann Defretin, et al. ultralytics/yolov5: v5. 0-yolov5-p6 1280 models, aws, supervise. ly and youtube integrations. *Zenodo*, 2021. 4

[19] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 2, 4

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2

[21] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[22] Minhyeok Lee, Sangwon Hwang, Chaewon Park, and Sangyoun Lee. Edgeconv with attention module for monocular depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 3

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017. 2

[24] Cheng-Jen Liu and Tsung-Nan Lin. Det: Depth-enhanced tracker to mitigate severe occlusion and homogeneous appearance problems for indoor multiple-object tracking. *IEEE Access*, 10:8287–8304, 2022. 3

[25] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 4

[26] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021. 5

[27] Armin Masoumian, Hatem A Rashwan, Julián Cristiano, M Salman Asif, and Domenec Puig. Monocular depth estimation using deep learning: A review. *Sensors*, 2022. 3

[28] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018. 2

[29] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 2, 6

[30] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[31] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1

[32] Taher Naderi, Amir Sadovnik, Jason Hayward, and Hairong Qi. Monocular depth estimation with adaptive geometric attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 3

[33] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks, 2019. 2

[34] Stephen Palmisano, Barbara Gillam, Donovan G Govan, Robert S Allison, and Julie M Harris. Stereoscopic perception of real depths at large distances. *Journal of vision*, 2010. 3

[35] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking, 2021. 2

[36] Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022. 3

[37] Yanlin Qian, Song Yan, Alan Lukežič, Matej Kristan, Joni-Kristian Kämäräinen, and Jiří Matas. Dal: A deep depth-aware long-term tracker. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7825–7832. IEEE, 2021. 3

[38] Kha Gia Quach, Huu Le, Pha Nguyen, Chi Nhan Duong, Tien Dai Bui, and Khoa Luu. Depth perspective-aware multiple object tracking, 2023. 3

[39] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[40] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 4

[41] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. 2

[42] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 4

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 2015. 2

[44] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 17–35. Springer, 2016. 6

[45] Francisco Romero-Ferrero, Mattia G Bergomi, Robert C Hinz, Francisco JH Heras, and Gonzalo G De Polavieja. Idtracker. ai: tracking all individuals in small or large collectives of unmarked animals. *Nature methods*, 16(2):179–182, 2019. 2

[46] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation, 2019. 2

[47] Tristan Walter and Iain D Couzin. Trex, a fast multi-animal tracking system with markerless identification, and 2d estimation of posture and visual fields. *Elife*, 10:e64000, 2021. 2

[48] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020. 4

[49] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 107–122. Springer, 2020. 6

[50] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. AB3DMOT: A baseline for 3d multi-object tracking and new evaluation metrics. *CoRR*, abs/2008.08063, 2020. 3

[51] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1, 2, 3, 4, 6, 7

[52] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 2

[53] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 3

[54] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[55] Libo Zhang, Junyuan Gao, Zhen Xiao, and Heng Fan. Animaltrack: A benchmark for multi-animal tracking in the wild. *International Journal of Computer Vision*, 131(2):496–513, 2023. 1, 2, 5, 6

[56] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2022. 2, 6

[57] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 6

[58] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pages 474–490. Springer, 2020. 1, 6

[59] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points, 2020. 2

[60] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021. 1

[61] Saman Zia, Buket Yuksel, Deniz Yuret, and Yucel Yemez. Rgb-d object recognition using deep convolutional neural networks. In *Proceedings of the IEEE International conference on computer vision workshops*, pages 896–903, 2017. 3